JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

CRICKET SCORE PREDICTION USING MACHINE LEARNING

[1] Sanket Aldar, [2] Kanishk Mane ,[3] Savita Sangam
[1,2,] Scholar, [3]Professor
Department of Information Technology,
SSJCOE, Dombivli, India

Abstract : Cricket is a widely-followed sport in India and around the world. In latest years, the T-20 version of this sport has grown in reputation. The Indian Premier League (IPL), a match primarily based totally on this system, has grown in reputation in the latest years. Cricket, on the alternative hand, is visible as a sport of chance. Predicting the winner of a match or healthy is likewise a fear amongst fanatics and followers. Meanwhile, technology is advancing at an astonishing pace. Following the schooling of a model, devices getting to know algorithms are used to expect something. As a result, in this study, we appoint a lot of supervised mastering strategies to are expecting the winners of Indian Premier League matches. Team names, match site, toss winner, toss decision, match winner, gained through what number of runs, and umpires' gift for the match are the various attributes of this system. Logistic regression, selection tree, random forest, SVM, naive Bayes, gradient boost, and KNN are one of the tasks intently monitored approaches.
.

Index Terms - Machine learning ,Regression models ,Time series analysis ,Ensemble methods ,Data preprocessing ,Feature selection ,Model evaluation, Deployment

I. INTRODUCTION

In cricket, especially in limited-overs formats like One Day Internationals (ODIs) and Twenty20 (T20) matches, predicting the final score can be crucial for teams to strategize effectively. Factors such as batting team, bowling team, venue conditions, current score, overs remaining, and wickets lost play significant roles in determining the outcome of a match. With the advent of machine learning techniques, predicting cricket scores has become an area of active research. This paper aims to contribute to this domain by proposing a machine learning-based approach to predict cricket scores.

Data analysis today, every data analyst needs to examine data sets and draw conclusions from the information to extract useful information from them. Data analytics techniques and algorithms are widely employed by the commercial industry to form accurate business decisions. Verification or refutation of experimental layouts, hypotheses and conclusions are also used by analysts and specialists. In recent years, analytics has been used to predict and draw various insights into the field of sports. Due to money, team spirit, loyalty to the city and the participation of a large number of fans, the results of the competitions are very important to all stakeholders (Jyothsna and Srikanth, 2019). Sports forecasting can be considered as one of the objectives of sports analysis, which aims to help decision makers to take advantage of competitors. Data analysis is especially common in sports. Cricket is accustomed to using the International Cricket Council (ICC) data analysis results. The barrier to this task depends on the collection of data historical data, the collection of data for future events, the knowledge required to interpret the collected data, and much more. The result of the game has become the center and concentration of the game (Naik et al., 2018). Most of the past research studies, did their research to develop an AI tool for ODI (One Day International) and test matches.

1.RELATED WORK

• HIGH LEVEL DESIGN

The architecture that will be utilized to produce software products is described in high level design. The architecture diagram depicts the overall system, defining the main components and their interfaces that will be produced for the product.

• ARCHITECTURE OF IPYTHON NOTEBOOK

The notebook takes interactive computing in a qualitatively new direction by providing a web-based application for capturing the entire computation process, including designing, documenting, and executing code, as well as conveying the results. The Jupyter notebook is formed using two parts: -A web application is a browser-based tool for interactive document writing that combines explanatory text, mathematics, computations, and the output rich media. -Notebook documents are a representation of all visible material in the web application, including calculation inputs and outputs, explanatory language, mathematics, graphics, and rich media object representations.

- **THE PYTHON KERNEL**

The Notebook, Qt console, I Python console in the terminal interfaces use the I Python Kernel, and third-party interfaces. The I I Python Kernel is a different process that runs user code and performs computing probable completions tasks. JSON messages transmitted across Zero MQ sockets are used by frontends like the notebook and the Qt console to interface with the I Python Kernel

.

- **THE NOTEBOOK**

The Notebook frontend performs a unique function. It not only runs your code, but it also saves it, together with any markdown notes, in an editable document called a notebook. When it is saved, the browser sends it to the notebook server as a JSON file with the. ipynb extension on disc. The figure depicts its functions briefly.

- **THE NOTEBOOK DESIGN**

For a designer or engineer to keep track of project progress from beginning to end, a design notebook is a good technique. During the design process, it's a location to keep track of research, observations, ideas, drawings, comments, and queries.

- **SUPERVISED MACHINE LEARNING**

Supervised gaining knowledge of is used withinside the sizable majority of real device gaining knowledge of applications. When you've got enter variables (x) and an output variable (Y), supervised gaining knowledge takes place whilst you follow a set of rules to examine the mapping feature from the enter to the output (Y = f) (X). The purpose is to estimate the mapping feature to the factor that you may forecast the output variables (Y) for brand spanking new enter data (x). Some of Examples of supervised machine learning techniques are Linear and logistic regression, multi-class classification, Decision Trees, and support vector machines. The data needed to train the algorithm for supervised learning must already be labeled with correct responses. For example, after being trained on a dataset of photos that are appropriately labelled with the animal's species and certain identifying traits, a classification algorithm will learn to identify animals. Regression and category problems are kind of supervised learning obligations. The motive of each demanding situations is to create an easy version which could expect the fee of the structured characteristic the usage of handiest the characteristic variables. The handiest distinction among the 2 obligations is that the structured function in regression is numerical, while in category it's miles categorical. The types of classification algorithms used in this project are:

- Logistic regression.
- Naive Bayes Classifier

. • K Nearest Neighbor.

- Support Vector Machines.

1.1.1 Population And Sample

matches or innings that you want to predict the scores for .This encompasses all cricket matches across various formats (Test, One Day Internationals, T20s) and leagues (international, domestic).For instance, if you're focusing on predicting scores in T20 matches, your population would consist of all T20 matches played worldwide .A sample is a subset of the population that is selected for analysis. It's impractical and often impossible to analyze the entire The population in your research refers to the entire group of interest, which in this case would be all cricket population, so researchers typically work with samples. In cricket score prediction research, your sample might consist of a subset of • Decision Trees.

matches or innings that you'll use to develop and test your prediction models .Ensure that your sample is representative of the population to ensure the generalizability of your findings. This means it should include a • Random Forest. diverse range of matches, teams, playing conditions, and other relevant factors.

1.1.2 Data and Source of Data

Official Cricket Boards and Organizations:

International Cricket Council (ICC): The governing body for international cricket, providing detailed statistics and match data for all international matches.

National Cricket Boards: Individual cricket boards of countries organize domestic matches and maintain databases of match statistics, which are often accessible on their official websites.

Cricket Statistics Websites:

ESPN Cric info: A comprehensive cricket website offering detailed statistics, scorecards, and match summaries for international and domestic matches.

Cricket Archive: A vast archive of cricket statistics covering matches from various eras, including international and domestic cricket.

How stat: Provides statistical analysis and player performance data for international cricket matches.

Commercial Data Providers:
Opta: Offers detailed sports data, including cricket statistics, for commercial use. It provides extensive data on player performance, match events, and contextual information.
Stats Perform: Provides sports data solutions for various sports, including cricket, offering detailed match statistics and performance analytics.

APIs and Databases:
Some websites and organizations offer APIs (Application Programming Interfaces) that allow developers to access cricket data programmatically. You can explore APIs provided by ICC, ESPN Cric info, or other cricket data providers**.** Databases such as Kaggle may also have cricket datasets contributed by researchers and enthusiasts.

2.PROPOSED METHOD

2.1 Obtain the dataset: .The dataset for analysis and prediction was obtained from www.kaggle.com, which included data from past IPL editions from 2008 to 2019. There were two datasets used in this study. On each ball of the match, the first gives us ballby-ball information from every match ever played in the IPL, including the batsman, bowler, runs, wicket, and more. The second dataset contains a summary of each match, including the teams involved, the winner, the toss winner, and other information for every match played in the IPL

2.2 Information Feature extraction and pre-processing:

The pre-processing stage cleans the dataset by deleting data that isn't necessary for obtaining results. During the pre-processing stage, data that has not been declared or tagged is eliminated. To extract the essential analysis, as well as for the prediction module, the data must be pre-processed and cleaned. The dataset was created using records from the last 11 years, or from season 2008 to 2019. Methods such as eliminating outliers, normalizing, and standardization are used to pre-process the data
.

2.3 Data format conversion:

Because a few of the dataset's attributes are categorical, classification is rather difficult. It may potentially have an impact on the model, resulting in incorrect predictions. Except for the target attribute (Winner), all categorical data in the dataset has been transformed to numeric format and standardized on a scale basis in this step. Ordinal encoding and one-hot encoding are the two most used approaches

2.4 Model Training:

The datasets were divided into two sections for training and testing before the model was trained. On one of the datasets, three regression models are used to predict the score: Lasso Regression, Ridge Regression, and Random Forest Regression. On the one hand, three predictive modeling classifiers, Support Vector Machine (SVM), Logistic Regressions are used for classification.

2.4.1 Random Forest: Random forest is an ensemble-based supervised learning methodology. Ensemble learning is a type of learning in which many decision trees are constructed and then combined to produce more accurate prediction models. The resulting of trees termed "Random Forest" is a mix of multiple decision trees. The random forest algorithm is not biassed because it is based on a majority vote and delivers the final prediction based on that voting. Random Forest and Decision Tree both employ the identical Equation(1) and Equation(2) formulas. The Random Forest method is based on the following principle: Phase 1: In this step, random rows from the training data set will be selected by assigning an arbitrary value. Phase 2: The decision tree is built based on the selection of random rows in this step. The output is then created from each decision tree. Phase 3: Using the frequency method, voting will be done on the created output. Phase 4: Based on the number of votes collected from the decision trees, the ultimate result is anticipated from step 3.

2.4.2 Ridge Regression: When the number of predictor variables in a set exceeds the number of observations, or when a data set exhibits multicollinearity, ridge regression is an approach to develop a parsimonious model (correlations between predictor variables). Ridge regression employs a ridge estimator, which is a sort of shrinkage estimator. Theoretically, shrinkage estimators generate new estimators that are closer to the "actual" population parameters. The ridge regression cost function is as follows: .5 Testing Data:

2.4.3 Lasso Regression: Lasso regression is a sort of linear regression that makes use of shrinkage. Data values are shrunk towards a central point, such as the mean, in shrinkage. The least absolute shrinkage and Selection Operator is an acronym that stands for Least Absolute Shrinkage and Selection Operator. Quadratic programming challenges, such as Lasso solutions, are best tackled with software (like Matlab). The intensity of the L1 penalty is controlled by a tuning parameter. is the amount of shrinkage in terms of: D = least-squares + lambda * summation (absolute values of the magnitude of the coefficients)

2.4.4 Support Vector Machine: Support vectors are data points that are closer to the hyperplane and have an impact on the hyperplane's location and orientation. We maximize the classifier's margin by using these support vectors. The goal of the SVM method is to maximize the distance between the data points and the hyperplane. Hinge loss is a loss function that aids in margin maximization.

2.4.5 Logistic Regression: Under the Supervised Learning approach, Logistic Regression is one of the most used Machine Learning algorithms. It's a method for predicting a categorical dependent variable from a set of independent variables. The steps below will be used to implement the Logistic Regression: Fitting Logistic Regression to the Training Set Predicting the Test Result (Creation of Confusion Matrix) Visualizing the Test Set Result

3. ALGORITHM

Step 1: Start

Step 2: Collect historical matches data and relevant factors.

Step 3: Preprocess the data by cleaning and encoding features.

Step 4: Engineer additional features like lag variables and external data integration.

Step 5: Select machine learning models such as linear regression or random forests.

Step 6: Train the selected models on the training dataset.

Step 7: Evaluate model performance using metrics like MAE or RMSE.

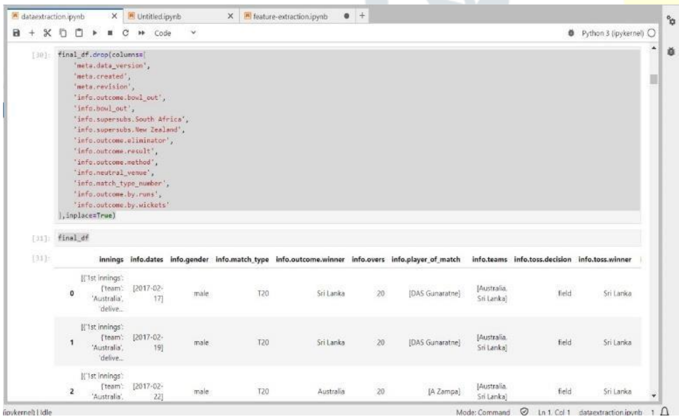
Step 8: Deploy the best-performing model for cricket score predicting

Step 9: Input new data for score predicting

Step 10: Monitor predicted score and update the model as needed.

Step 11: End.

4.RESULT.



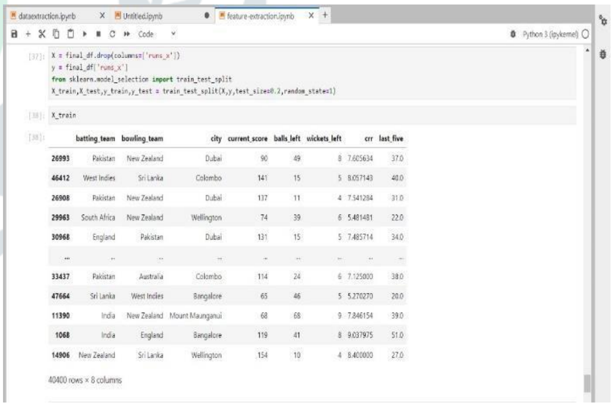
[30]:
```
final_df.drop(columns=[
    'meta.data_version',
    'meta.created',
    'meta.revision',
    'info.outcome.bowl_out',
    'info.bowl_out',
    'info.supersubs.South Africa',
    'info.supersubs.New Zealand',
    'info.outcome.eliminator',
    'info.outcome.result',
    'info.outcome.method',
    'info.neutral_venue',
    'info.match_type_number',
    'info.outcome.by.runs',
    'info.outcome.by.wickets'
],inplace=True)
```

[31]:
```
final_df
```

[31]:

	innings	info.dates	info.gender	info.match_type	info.outcome.winner	info.overs	info.player_of_match	info.teams	info.toss.decision	info.toss.winner
0	[{'1st innings': {'team': 'Australia', 'delive...	[2017-02-17]	male	T20	Sri Lanka	20	[DAS Gunaratne]	[Australia, Sri Lanka]	field	Sri Lanka
1	[{'1st innings': {'team': 'Australia', 'delive...	[2017-02-19]	male	T20	Sri Lanka	20	[DAS Gunaratne]	[Australia, Sri Lanka]	field	Sri Lanka
2	[{'1st innings': {'team': 'Australia',	[2017-02-22]	male	T20	Australia	20	[A Zampa]	[Australia, Sri Lanka]	field	Sri Lanka

FIG.4.1.DATA CLEANING.



[37]:
```
X = final_df.drop(columns=['runs_x'])
y = final_df['runs_x']
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.2,random_state=1)
```

[38]:
```
X_train
```

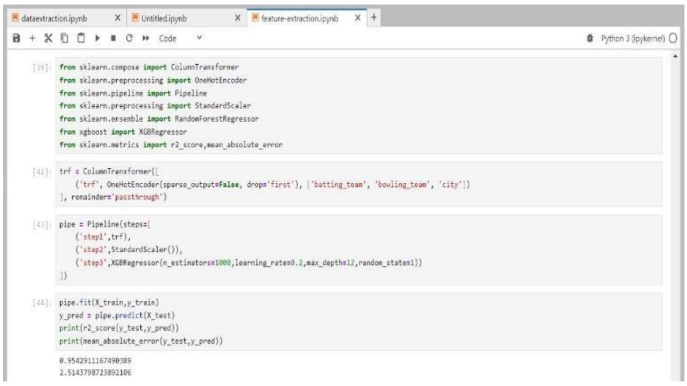
[38]:

	batting_team	bowling_team	city	current_score	balls_left	wickets_left	crr	last_five
26993	Pakistan	New Zealand	Dubai	90	49	8	7.605634	37.0
46412	West Indies	Sri Lanka	Colombo	141	15	5	8.057143	40.0
26908	Pakistan	New Zealand	Dubai	137	11	4	7.541284	31.0
29963	South Africa	New Zealand	Wellington	74	39	6	5.481481	22.0
30968	England	Pakistan	Dubai	131	15	5	7.485714	34.0
...	...	...	...	...	...	...	...	...
33437	Pakistan	Australia	Colombo	114	24	6	7.125000	38.0
47664	Sri Lanka	West Indies	Bangalore	65	46	5	5.270270	20.0
11390	India	New Zealand	Mount Maunganui	68	68	9	7.846154	39.0
1068	India	England	Bangalore	119	41	8	9.037975	51.0
14906	New Zealand	Sri Lanka	Wellington	154	10	4	8.400000	27.0

40400 rows × 8 columns

FIG.4.2.DATA TRAINING.

.



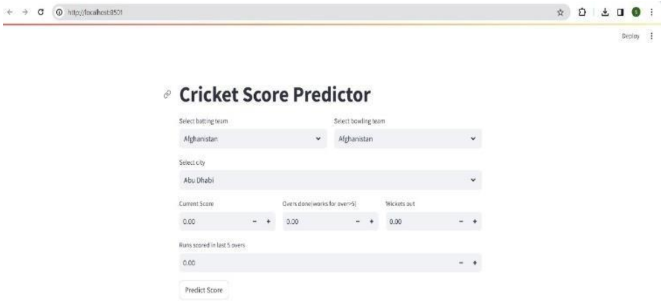
```
[39]: from sklearn.compose import ColumnTransformer
      from sklearn.preprocessing import OneHotEncoder
      from sklearn.pipeline import Pipeline
      from sklearn.preprocessing import StandardScaler
      from sklearn.ensemble import RandomForestRegressor
      from xgboost import XGBRegressor
      from sklearn.metrics import r2_score,mean_absolute_error

[42]: trf = ColumnTransformer([
          ('trf', OneHotEncoder(sparse_output=False, drop='first'), ['batting_team', 'bowling_team', 'city'])
      ], remainder='passthrough')

[43]: pipe = Pipeline(steps=[
          ('step1',trf),
          ('step2',StandardScaler()),
          ('step3',XGBRegressor(n_estimators=1000,learning_rate=0.2,max_depth=12,random_state=1))
      ])

[44]: pipe.fit(X_train,y_train)
      y_pred = pipe.predict(X_test)
      print(r2_score(y_test,y_pred))
      print(mean_absolute_error(y_test,y_pred))

      0.9542911167490389
      2.5143798723892106
```



Cricket Score Predictor

Select batting team: Afghanistan

Select bowling team: Afghanistan

Select city: Abu Dhabi

Current Score: 0.00

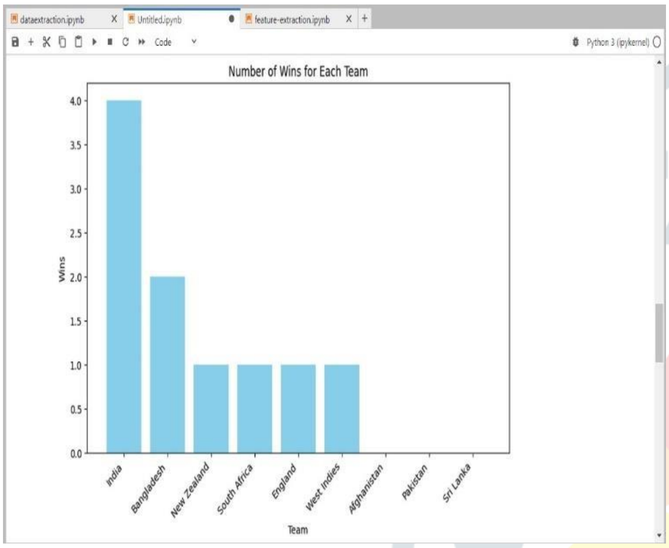
Overs done(works for over>5): 0.00

Wickets out: 0.00

Runs scored in last 5 overs: 0.00

FIG.4.3. MODEL DEVLOPMENT

FIG.4.4.HOMEPAGE



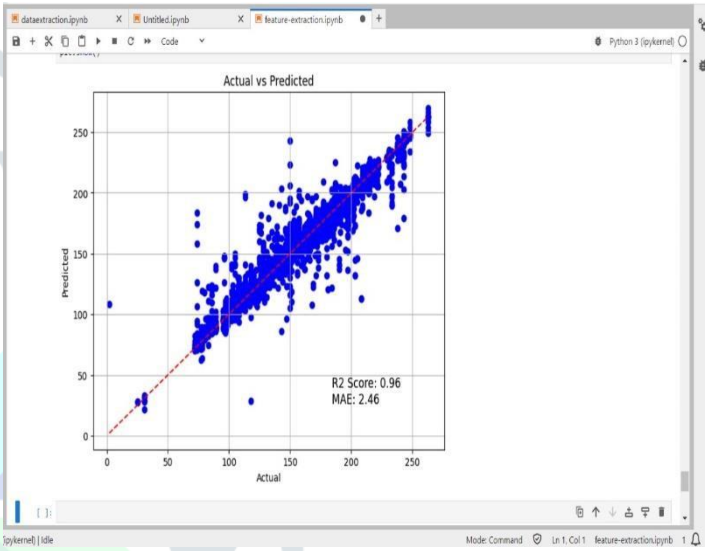


FIG.4.5.NO.OF WINS

FIG.4.6.PREDICTION

5.CONCLUSION

In this work, the data sets used have been collected from real T20 cricket matches and impractical features have been removed in pre processing with few other data cleaning steps. Additionally, suitable data is converted to a numeric form. First and fore most, the cleaned data which is used for win prediction is trained and classified with three classifiers SVM Linear, SVM RBF, Logistic Regression. Subsequently, the cleaned dataset which is used for runs prediction is trained with three regressors Random Forest, Ridge Regression, Lasso Regression, and python tool is used in both the predictions. Good results have been achieved using the SVM RBF classifier for a predict score and Random Forest Regressor for runs with an overall accuracy of 83% and 75% respectively. As our approach well predicts the T20 in the current scenario that is based on the historical records, it can be further extended after youngsters join the team, their history records are made available. Moreover, new season data can be added, and adding some new features like head-to-head win which are beneficial in increasing accuracy

6.FUTURE SCOPE.

This project proves that machine learning is extremely useful in predicting the scored of batting team in match. The algorithm used in this experiment has performed really well using the available attributes. The analysis done in our project can be useful for cricket enthusiasts and also help the team management to take better decisions. We have built 2 machine learning models using-Linear Regression and Random Forest Classifier. Random Forest Classifier is used to predict the winner based on toss and venue whereas Linear Regression is used to predict winner based on the target and runs scored by the chasing team in at least5 overs. As we know that T20 is a dynamic game, so we have done prediction of the winner considering the dynamic nature of the format. In the future we can consider Player's Performance as one of the attributes as we know that player's form is also one of the most important factor for a team to win a match. Accuracy of a model would increase if we consider player's performance

7. REFERENCES

[1] Chen, T., & Guestrin, C. (2016). XG Boost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

[2] Singh, A., & Kaur, H. (2020). Predicting One Day International Cricket Matches Using Machine Learning Techniques. International Journal of Computer Applications, 174(32), 13-17.

[3] Patel, P., & Patel, D. (2018). Prediction of Cricket Match Outcome Using Machine Learning Algorithms. In 2018 International Conference on Recent Trends in Image Processing and Pattern Recognition (RTIP2R) (pp. 1-6).

[4] https://youtu.be/tZd1okZiijo?si=_Iqaw_RCynuLu-6P.

[4] Rameshwari Lokhande, P. M. Chawan "Live Cricket Score and Winning Prediction" Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, Volume-5 | Issue1 , February 2018

[5]. Nimmag adda, Akhil, et al. "Cricket score and winning prediction using data mining." International Journal for Advance Research and Development 3.3 (2018): 299-302.